ON SOME USEFUL "INEFFICIENT" STATISTICS

By

FREDERICK MOSTELLER

Reprinted from ANNALS OF MATHEMATICAL STATISTICS
Vol. XVII, No. 4, December, 1946

Made in United States of America

1946

A DISSERTATION

PRESENTED TO THE

FACULTY OF PRINCETON UNIVERSITY

IN CANDIDACY FOR THE DEGREE

OF DOCTOR OF PHILOSOPHY

RECOMMENDED FOR ACCEPTANCE BY THE

DEPARTMENT OF

MATHEMATICS

JUNE, 1946

Reprinted from The Annals of Mathematical Statistics
Vol. XVII, No. 4, December, 1946

ON SOME USEFUL "INEFFICIENT" STATISTICS

By Frederick Mosteller
Princeton University

TABLE OF CONTENTS

PAGE
Summary 377
1. Introduction 377
2. Order statistics 379
3. Estimates of the mean of a normal distribution 386
4. Estimates of the standard deviation 389
4A. The range as an estimate of σ 390
4B. Quasi ranges for estimating σ 391
4C. The mean deviations about the mean and median 393
5. Estimation of the correlation coefficient 394
5A. Estimation of ρ when means and standard deviations are known 395
5B. Estimation of ρ when the parameters are unknown 399
5C. The use of averages for estimating ρ when the variance ratio is known 406
6. Acknowledgements 407

Summary. Several statistical techniques are proposed for economically analyzing large masses of data by means of punched-card equipment; most of these techniques require only a counting sorter. The methods proposed are designed especially for situations where data are inexpensive compared to the cost of analysis by means of statistically "efficient" or "most powerful" procedures. The principal technique is the use of functions of order statistics, which we call *systematic statistics*.

It is demonstrated that certain order statistics are asymptotically jointly distributed according to the normal multivariate law.

For large samples drawn from normally distributed variables we describe and give the efficiencies of rapid methods:

i) for estimating the mean by using 1, 2, $\cdots$, 10 suitably chosen order statistics; (cf. p. 386)

ii) for estimating the standard deviation by using 2, 4, or 8 suitably chosen order statistics; (cf. p. 389)

iii) for estimating the correlation coefficient whether other parameters of the normal bivariate distribution are known or not (three sorting and three counting operations are involved) (cf. p. 394).

The efficiencies of procedures ii) and iii) are compared with the efficiencies of other estimates which do not involve sums of squares or products.

1. Introduction. The purpose of this paper is to contribute some results concerning the use of order statistics in the statistical analysis of large masses of data. The present results deal particularly with estimation when normally distributed variables are present. Solutions to all problems considered have

been especially designed for use with punched-card equipment although for most of the results a counting sorter is adequate.

Until recently mathematical statisticians have spent a great deal of effort developing "efficient statistics" and "most powerful tests." This concentration of effort has often led to neglect of questions of economy. Indeed some may have confused the meaning of technical statistical terms "efficient" and "efficiency" with the layman's concept of their meaning. No matter how much energetic activity is put into analysis and computation, it seems reasonable to inquire whether the output of information is comparable in value to the input measured in dollars, man-hours, or otherwise. Alternatively we may inquire whether comparable results could have been obtained by smaller expenditures. In some fields where statistics is widely used, the collection of large masses of data is inexpensive compared to the cost of analysis. Often the value of the statistical information gleaned from the sample decreases rapidly as the time between collection of data and action on their interpretation increases. Under these conditions, it is important to have quick, inexpensive methods for analyzing data, because economy demands militate against the use of lengthy, costly (even if more precise) statistical methods. A good example of a practical alternative is given by the control chart method in the field of industrial quality control. The sample range rather than the sample standard deviation is used almost invariably in spite of its larger variance. One reason is that, after brief training, persons with slight arithmetical knowledge can compute the range quickly and accurately, while the more complicated formula for the sample standard deviation would create a permanent stumbling block. Largely as a result of simplifying and routinizing statistical methods, industry now handles large masses of data on production adequately and profitably. Although the sample standard deviation can give a statistically more efficient estimate of the population standard deviation, if collection of data is inexpensive compared to cost of analysis and users can compute a dozen ranges to one standard deviation, it is easy to see that economy lies with the less efficient statistic.

It should not be thought that inefficient statistics are being recommended for all situations. There are many cases where observations are very expensive, and obtaining a few more would entail great delay. Examples of this situation arise in agricultural experiments, where it often takes a season to get a set of observations, and where each observation is very expensive. In such cases the experimenters want to squeeze every drop of information out of their data. In these situations inefficient statistics would be uneconomical, and are not recommended.

A situation that often arises is that data are acquired in the natural course of administration of an organization. These data are filed away until the accumulation becomes mountainous. From time to time questions arise which can be answered by reference to the accumulated information. How much of these data will be used in the construction of say, estimates of parameters, depends on the precision desired for the answer. It will however often be less expensive to

get the desired precision by increasing the sample size by dipping deeper into the stock of data in the files, and using crude techniques of analysis, than to attain the required precision by restricting the sample size to the minimum necessary for use with "efficient" statistics.

It will often happen in other fields such as educational testing that it is less expensive to gather enough data to make the analysis by crude methods sufficiently precise, than to use the minimum sample sizes required by more refined methods. In some cases, as a result of the type of operation being carried out sample sizes are more than adequate for the purposes of estimation and testing significance. The experimenters have little interest in milking the last drop of information out of their data. Under these circumstances statistical workers would be glad to forsake the usual methods of analysis for rapid, inexpensive techniques that would offer adequate information, but for many problems such techniques are not available.

In the present paper several such techniques will be developed. For the most part we shall consider statistical methods which are applicable to estimating parameters. In a later paper we intend to consider some useful "inefficient" tests of significance.

2. Order statistics. If a sample $O_n = x_1', x_2', \cdots, x_n'$ of size n is drawn from a continuous probability density function $f(x)$, we may rearrange and renumber the observations within the sample so that

$$x_1 < x_2 < \cdots < x_n \tag{1}$$

(the occurrence of equalities is not considered because continuity implies zero probability for such events). The x_i's are sometimes called *order statistics*. On occasion we write $x(i)$ rather than x_i. Throughout this paper the use of primes on subscripted x's indicates that the observations are taken without regard to order, while unprimed subscripted x's indicate that the observations are order statistics satisfying (1). Similarly $x(n_i)$ will represent the n_ith order statistic, while $x'(n_i)$ would represent the n_ith observation, if the observations were numbered in some random order. The notation here is essentially the *opposite* of usual usage, in which attention is called to the order statistics by the device of primes or the introduction of a new letter. The present reversal of usage seems justified by the viewpoint of the article—that in the problems under consideration the use of order statistics is the natural procedure.

An example of a useful order statistic is the median; when $n = 2m + 1$ ($m = 0, 1, \cdots$), x_{m+1} is called the median and may be used to estimate the population median, i.e. u defined by

$$\int_{-\infty}^{u} f(t)\, dt = \tfrac{1}{2}.$$

In the case of symmetric distributions, the population mean coincides with u and x_{m+1} will be an unbiased estimate of it as well. When $n = 2m$ ($m = 1, 2,$

$\cdots$), the median is often defined as $\frac{1}{2}(x_m + x_{m+1})$. The median so defined is an unbiased estimate of the population median in the case of symmetric distributions; however for most asymmetric distributions $\frac{1}{2}(x_m + x_{m+1})$ will only be unbiased asymptotically, that is in the limit as n increases without bound. For another definition of the sample median see Jackson [8, 1921]. When x is distributed according to the normal distribution

$$N(x, a, \sigma^2) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-(1/2\sigma^2)(x-a)^2},$$

the variance of the median is well known to tend to $\pi\sigma^2/2n$ as n increases.

It is doubtful whether we can accurately credit anyone with the introduction of the median. However for some of the results in the theory of order statistics it is easier to give credit. In this section we will restrict the discussion to the order statistics themselves, as opposed to the general class of statistics, such as the range $(x_n - x_1)$, which are derived from order statistics. We shall call the general class of statistics which are derived from order statistics, and use the value ordering (1) in their construction, *systematic statistics*.

The large sample distribution of extreme values (examples x_r, x_{n-s+1} for r, s fixed and $n \to \infty$) has been considered by Tippett [17, 1925] in connection with the range of samples drawn from normal populations; by Fisher and Tippett [3, 1928] in an attempt to close the gap between the limiting form of the distribution and results tabled by Tippett [17], by Gumbel [5, 1934] (and in many other papers, a large bibliography is available in [6, Gumbel 1939]), who dealt with the more general case $r \geq 1$, while the others mentioned considered the special case of $r = 1$; and by Smirnoff who considers the general case of x_r, in [15, 1935] and also [16] the limiting form of the joint distribution of x_r, x_s, for r and s fixed as $n \to \infty$.

In the present paper we shall not usually be concerned with the distribution of extreme values, but shall rather be considering the limiting form of the joint distribution of $x(n_1)$, $x(n_2)$, $\cdots$, $x(n_k)$, satisfying

CONDITION 1. $\lim_{n\to\infty} \frac{n_i}{n} = \lambda_i$; $\quad i = 1, 2, \cdots, k$;

$$\lambda_1 < \lambda_2 < \cdots < \lambda_k .$$

In other words the proportion of observations less than or equal to $x(n_i)$ tends to a fixed proportion which is bounded away from 0 and 1 as n increases. K. Pearson [13, 1920] supplies the information necessary to obtain the limiting distribution of $x(n_1)$, and limiting joint distribution of $x(n_1)$, $x(n_2)$. Smirnoff gives more rigorous derivations of the limiting form of the marginal distribution of the $x(n_i)$ [15, 1935] and the limiting form of the joint distribution of $x(n_i)$ and $x(n_j)$ [16] under rather general conditions. Kendall [10, 1943, pp. 211–14] gives a demonstration leading to the limiting form of the joint distribution.

Since we will be concerned with statements about the asymptotic properties of the distributions of certain statistics, it may be useful to include a short dis-

cussion of their implications both practical and theoretical. If we have a statistic $\hat{\theta}(O_n)$ based on a sample O_n: x_1', x_2', $\cdots$, x_n' drawn from a population with cumulative distribution function $F(x)$ it often happens that the function $(\hat{\theta} - \theta)/\sigma_n = y_n$, where σ_n is a function of n is such that

$$\text{(A)} \qquad \lim_{n\to\infty} P(y_n < t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-\frac{1}{2}x^2}\, dx.$$

When this condition (A) is satisfied we often say: $\hat{\theta}$ *is asymptotically normally distributed with mean* θ *and variance* σ_n^2. We will not be in error if we use the statement in italics provided we interpret it as synonymous with (A). However there are some pitfalls which must be avoided. In the first place condition (A) may be true even if the distribution function of y_n, or of $\hat{\theta}$, has no moments even of fractional orders for any n. Consequently we do not imply by the italicized statement that $\lim_{n\to\infty} E[\hat{\theta}(O_n)] = \theta$, nor that $\lim_{n\to\infty} \{[E(\hat{\theta}^2) - [E(\hat{\theta})]^2\} = \sigma_n^2$, for, as mentioned, these expressions need not exist for (A) to be true. Indeed we shall demonstrate that Condition (A) is satisfied for certain statistics even if their distribution functions are as momentless as the startling distributions constructed by Brown and Tukey [1, 1946]. Of course it may be the case that all moments of the distribution of $\hat{\theta}$ exist and converge as $n \to \infty$ to the moments of a normal distribution with mean θ and variance σ_n^2. Since this implies (A), but not conversely, this is a stronger convergence condition than (A). (See for example J. H. Curtiss [2, 1942].) However the important implication of (A) is that for sufficiently large n each percentage point of the distribution of $\hat{\theta}$ will be as close as we please to the value which we would compute from a normal distribution with mean θ and variance σ_n^2, independent of whether the distribution of $\hat{\theta}$ has these moments or not.

Similarly if we have several statistics $\hat{\theta}_1$, $\hat{\theta}_2$, $\cdots$, $\hat{\theta}_k$, each depending upon the sample $O_n : x_1', x_2', \cdots, x_n'$, we shall say *that the* $\hat{\theta}_i$ *are asymptotically jointly normally distributed with means* θ_i, *variances* $\sigma_i^2(n)$, *and covariances* $\rho_{ij}\sigma_i\sigma_j$, when

$$\text{(B)} \qquad \lim_{n\to\infty} P(y_1 < t_1, y_2 < t_2, \cdots, y_k < t_k) = K \int_{-\infty}^{t_1} \int_{-\infty}^{t_2} \cdots \int_{-\infty}^{t_k} e^{-\frac{1}{2}Q^2}\, dx_1\, dx_2 \cdots dx_k,$$

where $y_i = (\hat{\theta}_i - \theta_i)/\sigma_i$, and Q^2 is the quadratic form associated with a set of k jointly normally distributed variables with variances unity and covariances ρ_{ij}, and K is a normalizing constant. Once again the statistics $\hat{\theta}_i$ may not have moments or product moments, the point that interests us is that the probability that the point with coordinates $(\hat{\theta}_1, \hat{\theta}_2, \cdots, \hat{\theta}_k)$ falls in a certain region in a k-dimensional space can be given as accurately as we please for sufficiently large samples by the right side of (B).

Since the practicing statistician is very often really interested in the probability that a point will fall in a particular region, rather than in the variance

or standard deviation of the distribution itself, the concepts of asymptotic normality given in (A) and (B) will usually not have unfortunate consequences. For example, the practicing statistician will usually be grateful that the sample size can be made sufficiently large that the probability of a statistic falling into a certain small interval can be made as near unity as he pleases, and will not usually be concerned with the fact that, say, the variance of the statistic may be unbounded.

Of course, a very real question may arise: how large must n be so that the probability of a statistic falling within a particular interval can be sufficiently closely approximated by the asymptotic formulas? If in any particular case the sample size must be ridiculously large, asymptotic theory loses much of its practical value. However for statistics of the type we shall usually discuss, computation has indicated that in many cases the asymptotic theory holds very well for quite small samples.

For the demonstration of the joint asymptotic normality of several order statistics we shall use the following two lemmas.

LEMMA 1. *If a random variable $\hat{\theta}(O_n)$ is asymptotically normally distributed converging stochastically to θ, and has asymptotic variance $\sigma^2(n) \underset{n\to\infty}{\to} 0$, where n is the size of the sample $O_n : x_1', x_2', \cdots, x_n'$, drawn from the probability density function $h(x)$, and $g(\hat{\theta})$ is a single-valued function with a nonvanishing continuous derivative $g'(\hat{\theta})$ in the neighborhood of $\hat{\theta} = \theta$, then $g(\hat{\theta})$ is asymptotically normally distributed converging stochastically to $g(\theta)$ with asymptotic variance $\sigma_n^2[g'(\theta)]^2$.*

PROOF. By the conditions of the lemma

$$\lim_{n\to\infty} P\left[\frac{\hat{\theta} - \theta}{\sigma_n} < t\right] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-\frac{1}{2}u^2}\, du.$$

Now if $t\sigma_n = \Delta\theta$, $\Delta\theta = \hat{\theta} - \theta$, using the mean value theorem there is a θ_1 in the interval $[\theta, \hat{\theta}]$, such that

$$g(\hat{\theta}) = g(\theta) + (\hat{\theta} - \theta)g'(\theta_1),$$

which implies

$$\lim_{n\to\infty} P\left(\frac{\hat{\theta} - \theta}{\sigma_n} < t\right) = \lim_{n\to\infty} P\left(\frac{g(\hat{\theta}) - g(\theta)}{\sigma_n g'(\theta_1)} < t\right), \qquad g'(\theta_1) \neq 0,$$

where θ_1 is a function of n. However $\lim_{\Delta\theta\to 0} g'(\theta_1) = g'(\theta)$ so we may write

$$\lim_{n\to\infty} P\left(\frac{\hat{\theta} - \theta}{\sigma_n} < t\right) = \lim_{n\to\infty} P\left(\frac{g(\hat{\theta}) - g(\theta)}{\sigma_n g'(\theta)} < t\right), \qquad g'(\theta) \neq 0.$$

where the form of the expression on the right is the one required to complete the proof of the lemma.

Of course if we have several random variables $\hat{\theta}_1, \hat{\theta}_2, \cdots, \hat{\theta}_k$, we can prove by an almost identical argument that

LEMMA 2. *If the random variables $\hat{\theta}_i(O_n)$ are asymptotically jointly normally*

distributed converging stochastically to θ_i, *and have asymptotic variances* $\sigma_i^2(n) \underset{n\to\infty}{\longrightarrow} 0$, *and covariances* $\rho_{ij}\sigma_i\sigma_j$, *where* n *is the size of the sample* $O_n : x_1', x_2', \cdots, x_n'$ *drawn from the probability density function* $h(x)$, *and* $g_i(\hat{\theta}_i)$, $i = 1, 2, \cdots, k$, *are single-valued functions with nonvanishing continuous derivatives* $g_i'(\hat{\theta}_i)$ *in the neighborhood of* $\hat{\theta}_i = \theta_i$, *then the* $g_i(\hat{\theta}_i)$ *are jointly asymptotically normally distributed with means* $g_i(\theta_i)$, *variances* $\sigma_i^2[g_i'(\theta_i)]^2$ *and covariances* $\rho_{ij}\sigma_i\sigma_j g_i'(\theta_i)g_j'(\theta_j)$.

The following condition represents restrictions on the probability density function $f(x)$ sufficient for the derivation of the limiting form of the joint distribution of the $x(n_i)$ satisfying Condition 1.

CONDITION 2. *The probability density function* $f(x)$ *is continuous, and does not vanish in the neighborhood of* u_i, *where*

$$\int_{-\infty}^{u_i} f(x)\,dx = \lambda_i, \qquad i = 1, 2, \cdots, k.$$

If we recall the discussion of condition (B) above, the theorem of Pearson and Smirnoff may be stated:

THEOREM 1. *If a sample* $O_n : x_1, x_2, \cdots, x_n$ *is drawn from* $f(x)$ *satisfying Condition 2, and if* $x(n_1)$, $x(n_2)$ *satisfy Condition 1 as* $n \to \infty$, *then* $x(n_1)$, $x(n_2)$ *are asymptotically distributed according to the normal bivariate distribution with means* u_1, u_2,

$$\int_{-\infty}^{u_i} f(x)\,dx = \lambda_i,$$

and variances

$$\sigma_i^2 = \frac{\lambda_i(1-\lambda_i)}{n[f(u_i)]^2}, \qquad i = 1, 2,$$

and covariance

$$\rho_{12}\sigma_1\sigma_2 = \frac{\lambda_1(1-\lambda_2)}{nf(u_1)f(u_2)}.$$

Theorem 1 has an obvious generalization which seems not to have been carried out in the literature. The generalization may be stated:

THEOREM 2. *If a sample* $O_n : x_1, x_2, \cdots, x_n$ *is drawn from* $f(x)$ *satisfying Condition 2, and if* $x(n_1), x(n_2), \cdots, x(n_k)$ *satisfy Condition 1 as* $n \to \infty$, *then the* $x(n_i)$, $i = 1, 2, \cdots, k$, *are asymptotically distributed according to the normal multivariate distribution, with means* u_i,

$$\int_{-\infty}^{u_i} f(x)\,dx = \lambda_i,$$

and variances

$$\sigma_i^2 = \frac{\lambda_i(1-\lambda_i)}{nf(u_i)^2}, \qquad i = 1, 2, \cdots, k,$$

and covariances

$$\rho_{ij}\sigma_i\sigma_j = \frac{\lambda_i(1-\lambda_j)}{nf(u_i)f(u_j)}, \qquad 1 \le i < j \le k.$$

PROOF. We shall carry out the demonstration for the uniform distribution

$$f(x) = \begin{cases} 1, & 0 \le x \le 1, \\ 0, & \text{elsewhere,} \end{cases}$$

and then utilize the fact that by a suitable transformation of the uniform distribution we may get any $f(x)$ satisfying Condition 2. Of course for the particular case of the uniform distribution all moments of the $x(n_i)$ exist and converge to those of the asymptotic theory.

The joint probability density of the $x(n_i)$, satisfying Condition 1 and drawn from $f(x)$, is given by

$$g[x(n_1), x(n_2), \cdots, x(n_k)] = \frac{n!}{(n_1-1)!(n-n_k)!\prod_{i=2}^{k}(n_i-n_{i-1}-1)!} \left(\int_0^{x(n_1)} dt_1\right)^{n_1-1}\left(\int_{x(n_k)}^{1} dt_{k+1}\right)^{n-n_k}\prod_{i=2}^{k}\left[\int_{x(n_{i-1})}^{x(n_i)} dt_i\right]^{n_i-n_{i-1}-1}. \tag{2}$$

Performing the indicated integrations we get from the right of (2)

$$Cx(n_1)^{n_1-1}\prod_{i=2}^{k}[x(n_i)-x(n_{i-1})]^{n_i-n_{i-1}-1}[1-x(n_k)]^{n-n_k}, \tag{3}$$

where C is the multinomial coefficient on the right of (2). It is well known that for the uniform distribution $E[x(n_i)] = \frac{n_i}{n+1}$, or asymptotically $\frac{n_i}{n}$, $i = 1, 2, \cdots, k$. We make the transformation $y_i = \left(x(n_i) - \frac{n_i}{n}\right)\sqrt{n}$, leading to

$$C_1\left(\frac{n_1}{n}+\frac{y_1}{\sqrt{n}}\right)^{n_1-1}\prod_{i=2}^{k}\left(\frac{n_i-n_{i-1}}{n}+\frac{[y_i-y_{i-1}]}{\sqrt{n}}\right)^{n_i-n_{i-1}-1} \cdot\left(\frac{n-n_k}{n}-\frac{y_k}{\sqrt{n}}\right)^{n-n_k}. \tag{4}$$

Using the usual technique of factoring out expressions like

$$\left(\frac{n_i-n_{i-1}}{n}\right)^{n_i-n_{i-1}-1},$$

we rewrite (4) with C_2 as a new constant, and setting $\lambda_i = \frac{n_i}{n}$

$$C_2\left(1+\frac{y_1}{\lambda_1\sqrt{n}}\right)^{n_1-1} \cdot\prod_{i=2}^{k}\left(1+\frac{(y_i-y_{i-1})}{(\lambda_i-\lambda_{i-1})\sqrt{n}}\right)^{n_i-n_{i-1}-1}\left(1-\frac{y_k}{(1-\lambda_k)\sqrt{n}}\right)^{n-n} . \tag{5}$$

Now taking the logarithm of (5), expanding, neglecting terms $O\left(\frac{1}{\sqrt{n}}\right)$ and higher, collecting terms and taking the antilogarithm we get the approximate asymptotic distribution of the order statistics

$$g(x(n_1), x(n_2), \cdots, x(n_k))$$

$$= C_3 \exp\left[-\frac{1}{2}\left\{\sum_{i=1}^{k} y_i^2 \frac{\lambda_{i+1} - \lambda_{i-1}}{(\lambda_{i+1} - \lambda_i)(\lambda_i - \lambda_{i-1})} - 2\sum_{i=2}^{k} \frac{y_i y_{i-1}}{\lambda_i - \lambda_{i-1}}\right\}\right], \tag{6}$$

where $\lambda_0 = 0$, $\lambda_{k+1} = 1$. Now setting up the matrix of the coefficients of the quadratic expression in the exponent

$$A_{ii} = \frac{\lambda_{i+1} - \lambda_{i-1}}{(\lambda_{i+1} - \lambda_i)(\lambda_i - \lambda_{i-1})}; \qquad A_{i,i-1} = A_{i-1,i} = -\frac{1}{\lambda_i - \lambda_{i-1}},$$

$i = 1, 2, \cdots, k$; $A_{ij} = 0, |i - j| > 1$. To obtain the variances and covariances we need

$$A^{ij} = \frac{\text{cofactor of } A_{ij} \text{ in } ||A_{ij}||}{\text{determinant } A_{ij}}$$

(see for example Wilks [18, p. 63 et seq.]). Now

$$|A| = \text{determinant } A_{ij} = \prod_{1}^{k+1} \frac{1}{\lambda_i - \lambda_{i-1}}; \tag{7}$$

$$\text{cofactor of } A_{ii} = \lambda_i(1 - \lambda_i)|A|, \quad i = 1, 2, \cdots, k.$$

$$\text{cofactor of } A_{ij} = \begin{cases} \lambda_i(1 - \lambda_j)|A|, & i < j \\ \lambda_j(1 - \lambda_i)|A|, & j < i. \end{cases}$$

This completes the proof for the uniform distribution.

If the uniform distribution is transformed into a probability density function $f(x)$ satisfying Condition 2, by an order preserving transformation, we appeal to Lemma 2. We notice that the $x(n_i)$ are transformed into $g[x(n_i)]$, and that the probability that $x(n_i)$ falls in the interval $[u_i, u_i + \Delta u_i]$ is transformed into the probability that $g[x(n_i)]$ falls in the interval $[g(u_i), g(u_i + \Delta u_i)]$. Using the mean value theorem we may write

$$g(u_i + \Delta u_i) = g(u_i) + \Delta u_i g'(u_i'),$$

where u_i' lies in the interval $[u_i, u_i + \Delta u_i]$. However

$$\lim_{\Delta u_i \to 0} g'(u_i') = g'(u_i).$$

The density for the uniform distribution in the interval $[u_i, u_i + \Delta u_i]$ is just Δu_i, and this same density will tend to $f(u_i)\Delta u_i g'(u_i)$. Therefore $g'(u_i) = 1/f(u_i)$, which completes the proof of Theorem 2.

It would often be useful to know the small sample distribution of the order statistics, particularly in the case where the sample is drawn from a normal.

Fisher and Yates' tables [4] give the expected values of the order statistics up to samples of size 50. However it would be very useful in the development of certain small sample statistics to have further information. It is perhaps too much to expect tabulated distribution functions, but at least the variances and covariances would be useful. A joint effort has resulted in the calculation for samples $n = 2, 3, \cdots, 10$ of the expected values to five decimal places, the variances to four decimal places, and the covariances to nearly two decimal places. It is expected that these tables will be published shortly.

3. Estimates of the mean of a normal distribution. It will be important in what follows to define efficiency and to indicate its interpretation. Then we shall construct some estimates of the means of certain distributions and compute their efficiencies. Except for the tables given, the discussion is applicable to the estimation of the mean of any symmetric distribution; and, of course, the concept of efficiency is still more general in its application. A statistic $\hat{\theta}(O_n)$, where O_n is the sample, is said to be an *efficient* estimate of θ if

i) $\sqrt{n}\,(\hat{\theta} - \theta)$ is asymptotically normally distributed with zero mean and finite variance, $\sigma^2(\hat{\theta})$, and

ii) for any other statistic $\hat{\theta}'$ with $\sqrt{n}(\hat{\theta}' - \theta)$ asymptotically normally distributed with zero mean and variance $\sigma^2(\hat{\theta}')$, $\sigma^2(\hat{\theta}) \leq \sigma^2(\hat{\theta}')$.

The ratio $\sigma^2(\hat{\theta})/\sigma^2(\hat{\theta}')$ is termed the efficiency of $\hat{\theta}'$ if $\hat{\theta}$ is an efficient estimate of θ. For discussion see Wilks [18, 1943]. The concepts of efficient statistic or estimate and of efficiency were introduced by R. A. Fisher. They serve as one measure of the amount of information a statistic draws from a sample. It is also common practice to speak of relative efficiencies, for example, of the statistics $\hat{\theta}'$ and $\hat{\theta}''$ described in ii) above, we say if $\sigma^2(\hat{\theta}') < \sigma^2(\hat{\theta}'')$ that the efficiency of $\hat{\theta}''$ relative to $\hat{\theta}'$ is the ratio of the smaller variance to the larger. This concept of efficiency has sometimes been used when the normality assumption has been violated by one or both statistics, when one or both are biased, and when small samples are considered. When used under these conditions the concept of efficiency becomes more difficult to interpret, although a comparison of the variation of two statistics about the value they are commonly estimating is often of value.

In the case of estimates of the mean a of a variable which is normally distributed according to $N(x, a, \sigma^2)$ from a sample of n, we can often express the variance of an asymptotically unbiased estimate as $\sigma^2(\hat{\theta}_i) = k_i\sigma^2/n$. The sample mean $\hat{\theta} = \Sigma x_i'/n$ is an efficient estimate of a with variance σ^2/n. Then in such cases the efficiency of $\hat{\theta}_i$ in estimating a is $1/k_i$. The interpretation is merely that to obtain the same precision using $\hat{\theta}_i$ as is possible with $\hat{\theta}$, one must use a sample k_i times as large.

Bearing in mind that we are at present searching for economical methods for analyzing large samples, it is clear that the concept of efficiency offers us a practical way of comparing cost of information with cost of obtaining it.

In the present section and in sections 4 and 5 we shall develop certain systematic estimates of parameters of normally distributed variables. Our procedure then will be to compare the efficiency of the systematic estimates with the efficient statistic for estimating the parameter in question, and also in sections 4 and 5 we compare our estimates with a statistic not involving squares or products. Of course the efficient statistic for estimating the mean of a normal is the sample mean, therefore in this section we will only compare our estimates with the sample mean.

We can construct unbiased estimates of the mean of a normal distribution from linear combinations of suitably chosen order statistics. These systematic statistics will be asymptotically normally distributed if the order statistics from which they are derived satisfy Condition 1. We will restrict ourselves to a useful practical case where equal weights are used. In other words the estimate discussed is just the average of k order statistics $k^{-1}\Sigma x(n_i)$. Suppose $x(n_i)$, $i = 1, 2, \cdots, k$ satisfy Condition 1, that $E[x(n_i)] = E[x(n_{k-i+1})]$, so that $E[\Sigma x(n_i)] = a$. An important unsolved question is to discover what spacing of the $x(n_i)$ will yield minimum variance, and thereafter at what rate does the efficiency of this optimumly spaced estimate increase with k. Computational methods bog down rapidly after $k = 3$. Because so little is known about this problem it seems worthwhile to offer some results for three arbitrary spacings (these results are of course useful in analyzing data).

If the $x(n_i)$ satisfy Theorem 2 we may approximate the variance of the systematic statistic $\hat{\theta}_k = \Sigma x(n_i)/k$ by the usual formula

$$\sigma^2(\hat{\theta}_k) = E[\Sigma x(n_i)/k]^2 - [E(\Sigma x(n_i)/k)]^2. \tag{8}$$

We lose no generality by assuming the mean and variance of the underlying normal to be 0 and 1 respectively. Then using the fact that $\Sigma u_i = 0$, and the result of Theorem 1 we rewrite (8) as

$$\sigma^2(\hat{\theta}_k) = E[\Sigma(x(n_i) - u_i)/k]^2 = \frac{1}{k^2 n}\left[\sum_{i=1}^{k} \frac{\lambda_i(1-\lambda_i)}{f_i^2} + 2\sum_{i<j} \frac{\lambda_i(1-\lambda_j)}{f_i f_j}\right], \tag{9}$$

where $f_m = f(u_m)$.

Using the symmetry which makes $\lambda_i = 1 - \lambda_{k-i+1}$, $f_i = f_{k-i+1}$, and the fact that for $k = 2r + 1$, $f_{r+1} = 1/\sqrt{2\pi}$, $\lambda_{r+1} = \frac{1}{2}$, we may simplify the right side of equation (9) with the following results for $k = 1, 2, \cdots, 7$. The factor $1/k^2$ has not been disturbed. We also write the general formulas for the simplified form of (9), but we omit a rather lengthy combinatorial argument which establishes the generalization.

$$k = 1\text{:}\ \frac{\pi}{2n}$$

$$k = 2\text{:}\ \frac{2\lambda_1}{4nf_1^2}$$

(10)

$$k = 3\text{: } \frac{2}{9n}\left[\frac{\lambda_1}{f_1^2} + \frac{\lambda_1\sqrt{2\pi}}{f_1} + \frac{\pi}{4}\right]$$

$$k = 4\text{: } \frac{2}{16n}\left[\frac{\lambda_1}{f_1^2} + \frac{2\lambda_1}{f_1 f_2} + \frac{\lambda_2}{f_2^2}\right]$$

$$k = 5\text{: } \frac{2}{25n}\left[\frac{\lambda_1}{f_1^2} + \frac{2\lambda_1}{f_1 f_2} + \frac{\lambda_2}{f_2^2} + \sqrt{2\pi}\left(\frac{\lambda_1}{f_1} + \frac{\lambda_2}{f_2}\right) + \frac{\pi}{4}\right]$$

$$k = 6\text{: } \frac{2}{36n}\left[\frac{\lambda_1}{f_1^2} + \frac{\lambda_2}{f_2^2} + \frac{\lambda_3}{f_3^2} + \frac{2\lambda_1}{f_1 f_2} + \frac{2\lambda_1}{f_1 f_3} + \frac{2\lambda_2}{f_2 f_3}\right]$$

$$k = 7\text{: } \frac{2}{49n}\left[\frac{\lambda_1}{f_1^2} + \frac{\lambda_2}{f_2^2} + \frac{\lambda_3}{f_3^2} + \frac{2\lambda_1}{f_1 f_2} + \frac{2\lambda_1}{f_1 f_3} + \frac{2\lambda_2}{f_2 f_3} + \sqrt{2\pi}\left(\frac{\lambda_1}{f_1} + \frac{\lambda_2}{f_2} + \frac{\lambda_3}{f_3}\right) + \frac{\pi}{4}\right]$$

$$k = 2r\text{: } \frac{2}{(2r)^2 n}\left[\sum_{i=1}^{r}\frac{\lambda_i}{f_i^2} + 2\sum_{1\le i<j\le r}\frac{\lambda_i}{f_i f_j}\right], \qquad r \ge 1$$

$$k = 2r+1\text{: } \frac{2}{(2r+1)^2 n}\left[\frac{(2r)^2 n}{2}\,\sigma^2(\hat{\theta}_{2r}) + \sqrt{2\pi}\sum_{i=1}^{r}\frac{\lambda_i}{f_i} + \frac{\pi}{4}\right], \qquad r \ge 1.$$

In addition to the possibility of minimizing the equations of (10) by numerical methods, three other procedures suggest themselves: i) to space the order statistics uniformly in probability; ii) to choose those k order statistics whose expected values are equal to the expected values of the order statistics in a sample of size k drawn from a unit normal; iii) to choose $\lambda_i = (i - \frac{1}{2})/k$. The following table lists for $k = 1$, 2, and 3 the expected values u_i of the order statistics and the probability to the left of the expected values λ_i for each of the procedures. The chosen order statistics are counted from left to right. It will be noticed that the third method gives very good results, and has the value of simplicity of formula. The following table gives a comparison between the efficiencies resulting from spacing by the three methods. The three optimum cases are included for completeness.

Statisticians planning to use the method of expected values suggested above will find Fisher and Yates [4, 1943] table of the expected values of the order statistics in samples of size k drawn from a unit normal helpful for computing the λ_i. Alternatively the following table of λ_i might be used.

As an example of the use of Table III, suppose we are using the expected value method for estimating the mean of a large sample drawn from a normal distribution $N(x, a, \sigma^2)$. If we are willing to use 6 observations out of 1000 for this purpose Table III indicates the selection of x_{103}, x_{261}, x_{421}, x_{580}, x_{740}, x_{898}. Furthermore Table II indicates that the variance of the estimate of a based on the average of these six observations will be approximately $\sigma^2/.948n$, $n = 1000$.

4. Estimates of the standard deviation. The statistic

$$s^2 = \sum_{i=1}^{n} (x_i' - \bar{x})^2/(n - 1),$$

TABLE I

Comparison of the order statistics which would be chosen according to each of the four procedures for subsamples of $k = 1, 2, 3$

k	Order Statistic	Optimum		Equal Probability		Expected Values		$\lambda_i=(i-\frac{1}{3})/k$	
		u_i	λ_i	u_i	λ_i	u_i	λ_i	u_i	λ_i
1	First	.0000	.5000	.0000	.5000	.0000	.5000	.0000	.5000
2	First	−.6121	.2702	−.4307	.3333	−.5642	.2863	−.6745	.2500
	Second	.6121	.7298	.4307	.6667	.5642	.7137	.6745	.7500
3	First	−.9056	.1826	−.6745	.2500	−.8463	.1967	−.9674	.1667
	Second	.0000	.5000	.0000	.5000	.0000	.5000	.0000	.5000
	Third	.9056	.8174	.6745	.7500	.8463	.8033	.9674	.8333

TABLE II

Comparison of the efficiencies of four methods of spacing k order statistics used in the construction of an estimate of the mean

k	$\lambda_i=i/(k+1)$	Expected Values*	$\lambda_i=(i-\frac{1}{3})/k$	Optimum
1	.637	.637	.637	.637
2	.793	.809	.808	.810
3	.860	.878	.878	.879
4	.896	.914	.913	
5	.918	.933	.934	
6	.933	.948	.948	
7	.944	.956	.957	
8	.952	.963	.963	
9	.957	.968	.969	
10	.962	.972	.973	

* The u_i are chosen equal to the expected values of the order statistics of a sample of size k.

where $\bar{x} = \sum_{i=1}^{n} x_i'/n$ is well known to be an unbiased estimate of the population variance σ^2, for $n > 1$. However s is not in general an unbiased estimate of σ. We are not interested here in the question of when we should estimate σ and when it is more advantageous to estimate σ^2. All we want is to have an

unbiased estimate of σ, based on sums of squares, to compare with another unbiased estimate based on order statistics. In the case of observations drawn from a normal distribution

$$s' = \frac{(\frac{1}{2}n)^{\frac{1}{2}}\Gamma(\frac{1}{2}[n-1])}{\Gamma(\frac{1}{2}n)} \sqrt{\frac{\Sigma(x_i' - \bar{x})^2}{n}}, \tag{11}$$

is an unbiased estimate of σ (see for example Kenney [11], with variance

$$\sigma^2(s') = \left\{\frac{1}{2}\left[\frac{\Gamma(\frac{1}{2}[n-1])}{\Gamma(\frac{1}{2}n)}\right]^2 (n-1) - 1\right\}\sigma^2. \tag{12}$$

TABLE III*

$P(x < u_{i|k}) \times 10^4$, $u_{i|k} = E(x_{i|k})$, $x_{i|k}$ is the ith order statistic in a sample of size k drawn from a normal distribution $N(x, 0, 1)$

k \ i	1	2	3	4	5	6	7	8	9	10
1	5000									
2	2863	7137								
3	1987	5000	8013							
4	1516	3832	6168	8484						
5	1224	3103	5000	6897	8776					
6	1025	2605	4201	5799	7395	8975				
7	0881	2244	3622	5000	6378	7756	9119			
8	0773	1971	3182	4394	5606	6818	8030	9227		
9	0688	1756	2837	3919	5000	6082	7163	8244	9312	
10	0619	1584	2559	3536	4512	5488	6464	7441	7416	9381

* The table is given to more places than necessary for the purpose suggested because it may be of interest in other applications. The $E(x_{i|k})$ from which the table was derived were computed to five decimal places.

For most practical purposes however, when $n > 10$, the bias in s is negligible. For large samples $\sigma^2(s')$ approaches $\sigma^2/2n$.

4A. The range as an estimate of σ. As mentioned in the Introduction, section 1, it is now common practice in industry to estimate the standard deviation by means of a multiple of the range $R' = c_n(x_n - x_1)$, for small samples, where $c_n = 1/[E(y_n) - E(y_1)]$, y_n and y_1 being the greatest and least observations drawn from a sample of size n from a normal distribution $N(y, a, 1)$. Although we are principally interested in large sample statistics, for the sake of completeness, we shall include a few remarks about the use of the range in small samples.

Now R' is an unbiased estimate of σ, and its variance may be computed for small samples, see for example Hartley [7, 1942]. In the present case, although both R' and s' are unbiased estimates of σ, they are not normally distributed,

nor are we considering their asymptotic properties; therefore the previously defined concept of efficiency does not apply. We may however use the ratio of the variances as an arbitrary measure of the relative precision of the two statistics. The following table lists the ratio of the variances of the two statistics, as well as the variances themselves expressed as a multiple of the population variance for samples of size $n = 2, 3, \cdots, 10$.

4B. Quasi ranges for estimating σ. The fact that the ratio $\sigma^2(s')/\sigma^2(R')$ falls off in Table IV as n increases makes it reasonable to inquire whether it might not be worthwhile to change the systematic estimate slightly by using the statistic $c_{1|n}[x_{n-1} - x_2]$, or more generally $c_{r|n}[x_{n-r} - x_{r+1}]$ where $c_{r|n}$ is the multiplicative constant which makes the expression an unbiased estimate of σ (in particular $c_{r|n}$ is the constant to be used when we count in $r + 1$ observations from each end of a sample of size n, thus $c_{r|n} = 1/[E(y_{n-r} - y_{r+1})]$ where the

TABLE IV

Relative precision of s' and R', and their variances expressed as a multiple of σ^2, the population variance

n	$\sigma^2(s')/\sigma^2(R')$	$\sigma^2(s')/\sigma^2$	$\sigma^2(R')/\sigma^2$
2	1.000	.570	.570
3	.990	.273	.276
4	.977	.178	.182
5	.962	.132	.137
6	.932	.104	.112
7	.910	.0864	.0949
8	.889	.0738	.0830
9	.869	.0643	.0740
10	.851	.0570	.0670

y's are drawn from $N(y, a, 1)$). This is certainly the case for large values of n, but with the aid of the unpublished tables mentioned at the close of section 2, we can say that it seems not to be advantageous to use $c_{1|n}[x_{n-1} - x_2]$ for $n \leq 10$. Indeed the variance $c_{1|10}[x_9 - x_2]$, for the unit normal seems to be about .10, as compared with $\sigma^2(R')/\sigma^2 = .067$ as given by Table IV, for $n = 10$. The uncertainty in the above statements is due to a question of significant figures.

Considerations which suggest constructing a statistic based on the difference of two order statistics which are not extreme values in small samples, weigh even more heavily in large samples. A reasonable estimate of σ for normal distributions, which could be calculated rapidly by means of punched-card equipment is

$$\hat{\sigma} = \frac{1}{c}[x(n_2) - x(n_1)], \tag{13}$$

where the $x(n_i)$ satisfy Condition 1, and where $c = u_2 - u_1$, u_2 and u_1 are the expected values of the n_2 and n_1 order statistics of a sample of size n drawn from a unit normal. Without loss of generality we shall assume the x_i are drawn from a unit normal. Furthermore we let $\frac{n_2}{n} = \lambda_2 = 1 - \lambda_1 = 1 - \frac{n_1}{n}$. Of course σ will be asymptotically normally distributed, with variance

$$\sigma^2(\hat{\sigma}) = \frac{2}{nc^2}\left[\frac{\lambda_1(1-\lambda_1)}{[f(u_1)]^2} + \frac{\lambda_2(1-\lambda^2)}{[f(u_2)]^2} - \frac{2\lambda_1(1-\lambda_2)}{f(u_1)f(u_2)}\right]. \tag{14}$$

Because of symmetry $f(u_1) = f(u_2)$; using this and the fact that $\lambda_1 = 1 - \lambda_2$, we can reduce (14) to

$$\sigma^2(\hat{\sigma}) = \frac{2}{nc^2}\frac{\lambda_1(1-2\lambda_1)}{[f(u_1)]^2}. \tag{15}$$

We are interested in optimum spacing in the minimum variance sense. The minimum for $\sigma^2(\hat{\sigma})$ occurs when $\lambda_1 \doteq .0694$, and for that value of λ_1, $\sigma^2(\hat{\sigma}) \doteq .767\ \sigma^2/n$. Asymptotically s' is also normally distributed, with $\sigma^2(s') = \sigma^2/2n$. Therefore we may speak of the efficiency of $\hat{\sigma}$ as an estimate of σ as .652. It is useful to know that the graph of $\sigma^2(\hat{\sigma})$ is very flat in the neighborhood of the minimum, and therefore varying λ_1 by .01 or .02 will make little difference in the efficiency of the estimate $\hat{\sigma}$ (providing of course that c is appropriately adjusted). K. Pearson [13] suggested this estimate in 1920. It is amazing that with punched-card equipment available it is practically never used when the appropriate conditions described in the Introduction are present.

The occasionally used semi-interquartile range, defined by $\lambda_1 = .25$ has an efficiency of only .37 and an efficiency relative to $\hat{\sigma}$ of only .56.

As in the case of the estimate of the mean by systematic statistics, it is pertinent to inquire what advantage may be gained by using more order statistics in the construction of the estimate of σ. If we construct an estimate based on four order statistics, and then minimize the variance, it is clear that the extreme pair of observations will be pushed still further out into the tails of the distribution. This is unsatisfactory from two points of view in practice: i) we will not actually have an infinite number of observations, therefore the approximation concerning the normality of the order statistics may not be adequate if λ_1 is too small, even in the presence of truly normal data; ii) the distribution functions met in practice often do not satisfy the required assumption of normality, although over the central portion of the function containing most of the probability, say except for the 5% in each tail normality may be a good approximation. In view of these two points it seems preferable to change the question slightly and ask what advantage will accrue from holding two observations at the optimum values just discussed (say $\lambda_1 = .07$, $\lambda_2 = .93$) and introducing two additional observations more centrally located.

We define a new statistic

$$\hat{\sigma}' = \frac{1}{c'}[x(n_4) + x(n_3) - x(n_2) - x(n_1)], \tag{16}$$

$c' = E[x(n_4) + x(n_3) - x(n_2) - x(n_1)]$, where the observations are drawn from a unit normal. We take $\lambda_1 = 1 - \lambda_4$, $\lambda_2 = 1 - \lambda_3$, $\lambda_1 = .07$. It turns out that $\sigma^2(\hat{\sigma}')$ is minimized for λ_2 in the neighborhood of .20, and that the efficiency compared with s' is a little more than .75. Thus an increase of two observations in the construction of our estimate of σ increases the efficiency from .65 to .75. We get practically the same result for $.16 \leq \lambda_2 \leq .22$.

Furthermore, it turns out that using $\lambda_1 = .02$, $\lambda_2 = .08$, $\lambda_3 = .15$, $\lambda_4 = .25$, $\lambda_5 = .75$, $\lambda_6 = .85$, $\lambda_7 = .92$, $\lambda_8 = .98$, one can get an estimate of σ based on eight order statistics which has an efficiency of .896. This estimate is more efficient than either the mean deviation about the mean or median for estimating σ. The estimate is of course

$$\hat{\sigma}'' = [x(n_8) + x(n_7) + x(n_6) + x(n_5) - x(n_4) - x(n_3) - x(n_2) - x(n_1)]/C,$$

where $C = 10.34$.

To summarize: in estimating the standard deviation σ of a normal distribution from a large sample of size n, an unbiased estimate of σ is

$$\hat{\sigma} = \frac{1}{c}(x_{n-r+1} - x_r),$$

where $c = E(y_{n-r+1} - y_r)$ where the y's are drawn from $N(y, a, 1)$. The estimate $\hat{\sigma}$ is asymptotically normally distributed with variance

$$\sigma^2(\hat{\sigma}) = \frac{2}{nc^2}\frac{\lambda_1(1 - 2\lambda_1)}{[f(u_1)]^2},$$

where $\lambda_1 = r/n$, $f(u_1) = N(E(x_r), 0, \sigma^2)$. We minimize $\sigma^2(\hat{\sigma})$ for large samples when $\lambda_1 \doteq .0694$, and for that value of λ_1,

$$\sigma^2_{\text{opt}}(\hat{\sigma}) \doteq \frac{.767\sigma^2}{n}.$$

The unbiased estimate of σ

$$\hat{\sigma}' = \frac{1}{c'}(x_{n-r+1} + x_{n-s+1} - x_s - x_r)$$

may be used in lieu of $\hat{\sigma}$. If $\lambda_1 = r/n$, $\lambda_2 = s/n$ we find

$$\sigma^2(\hat{\sigma}' \mid \lambda_1 = .07, \lambda_2 = .20) \doteq \frac{.66\sigma^2}{n}.$$

4C. The mean deviations about the mean and median. The next level of computational difficulty we might consider for the construction of an estimate of σ is the process of addition. The mean deviation about the mean is a well known, but not often used statistic. It is defined by

$$\text{m.d.} = \sum_{i=1}^{n} | x_i' - \bar{x} | /n. \tag{17}$$

For large samples from a normal distribution the expected value of m.d. is $\sqrt{\frac{2}{\pi}}\sigma$, therefore to obtain an unbiased estimate of σ we define the new statistic $A = \sqrt{\frac{\pi}{2}}$ m.d. Now for large samples A has variance $\sigma^2[\frac{1}{2}(\pi - 2)]/n$, or an efficiency of .884. However there are slight awkwardnesses in the computation of A which the mean deviation about the median does not have.

It turns out that for samples of size $n = 2m + 1$ drawn from a normal distribution $N(y, a, 0)$ the statistic

$$M' = \sqrt{\frac{\pi}{2}} \frac{\sum |x_i - x_{m+1}|}{2m} \tag{18}$$

asymptotically has mean σ and variance

$$\sigma^2(M') = \frac{1}{2m}\left(\frac{\pi - 2}{2}\right)\sigma^2. \tag{19}$$

Thus in estimating the standard deviation of a normal distribution from large samples we can get an efficiency of .65 by the judicious selection of two observations from the sample, an efficiency of .75 by using four observations, and an efficiency of .88 by using the mean deviation of all the observations from either the mean or the median of the sample, and an efficiency of .90 by using eight order statistics.

5. Estimation of the correlation coefficient. In the present section we consider the estimation of the correlation coefficient of a normal bivariate population:

$$f(x, y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2(1-\rho^2)}\left(\frac{(x-a)^2}{\sigma_x^2} + \frac{(y-b)^2}{\sigma_y^2} - \frac{2\rho(x-a)(y-b)}{\sigma_x\sigma_y}\right)\right]. \tag{20}$$

The efficient estimate of ρ in a sample $O_n : (x_1', y_1'), (x_2', y_2'), \cdots, (x_n', y_n')$ drawn from the density (20) is

$$r = \frac{\sum(x_i' - \bar{x}')(y_i' - \bar{y})}{[\sum(x_i' - \bar{x})^2 \sum(y_i' - \bar{y})^2]^{\frac{1}{2}}}. \tag{21}$$

There are numerous other techniques in the literature for estimating ρ, among them i) the tetrachoric correlation coefficient which depends on a four-fold table, ii) the adjusted rank correlation coefficient which depends on assigning ranks to the x and y observations. These and other estimates of the correlation coefficient are discussed by Kendall [10].

We shall be concerned with the construction of some estimates of the correlation coefficient which are particularly adapted for use with punched-card equipment. A counting sorter is adequate for the first two cases discussed; in line with our previous development we shall then consider a technique which uses simple addition of the observed values, but does not require sums of squares or products (in the special case where variances of x and y are equal).

5A. Estimation of ρ when means and standard deviations are known. Let us suppose that the means and variances of the variables x and y, distributed according to (20) are given, and consider the problem of estimating the correlation coefficient ρ from a sample of size n. There will be no generality lost by assuming $a = b = 0$, $\sigma_x^2 = \sigma_y^2 = 1$. The technique used will be to construct lines $y = 0$, $x = \pm k$, which cut the xy-plane into six parts. We will form an estimate of ρ based upon the number of observations falling in the four corners. Figure 1 represents the lines laid out in the manner suggested in connection with a scatter diagram of 25 observations; naturally the method is recommended for

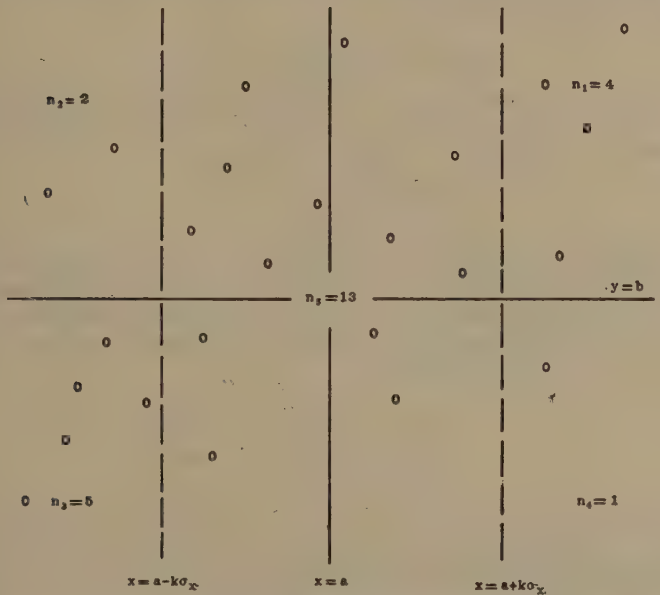

FIG. 1. DIAGRAM OF THE CONSTRUCTION DESCRIBED IN PARAGRAPH 5A WITH A SAMPLE OF 25 OBSERVATIONS SUPERIMPOSED

use only with large samples, the 25 observations are for purposes of illustration only. More specifically after assigning the special values mentioned immediately above to the means and variances in (20), we define

$$p_1 = \int_0^\infty \int_k^\infty f(x, y)\, dx\, dy, \qquad p_3 = \int_{-\infty}^0 \int_{-\infty}^{-k} f(x, y)\, dx\, dy,$$

(22)
$$p_2 = \int_0^\infty \int_{-\infty}^{-k} f(x, y)\, dx\, dy, \qquad p_4 = \int_{-\infty}^0 \int_k^\infty f(x, y)\, dx\, dy,$$

$$p_5 = \int_{-\infty}^\infty \int_{-k}^{k} f(x, y)\, dx\, dy = \int_{-k}^{k} N(x, 0, 1)\, dx.$$

We denote by n_i the number of observations falling into the region containing probability density p_i. Of course $\sum_{i=1}^{5} n_i = n$. Now we may write the joint probability distribution of the n_i as

$$g(n_1, n_2, n_3, n_4) = \frac{n!}{\prod_1^5 n_i!} \prod_{i=1}^{5} p_i^{n_i}. \tag{23}$$

remembering that $n_5 = n - \sum_1^4 n_i$.

We shall now derive the maximum likelihood estimate of ρ from (23). Taking the logarithm of (23) we have

$$\log g = \log c + \sum_{i=1}^{5} n_i \log p_i, \tag{24}$$

where c is the multinomial coefficient on the right of (23). Differentiating (24) with respect to ρ gives

$$\frac{d(\log g)}{d\rho} = \sum_{i=1}^{4} \frac{n_i \dot{p}_i}{p_i}. \tag{25}$$

where $\dot{p}_i = \frac{dp_i}{d\rho}$; of course $\frac{dp_5}{d\rho} = 0$ because p_5 is functionally independent of ρ. To get $\tilde{\rho}$, the maximum likelihood estimate of ρ, under our restrictions, we must equate the right of (25) to zero and solve for ρ. Before proceeding it will be useful to note the following relations:

$$p_1 = p_3 \; ; \; p_2 = p_4$$

$$\dot{p}_1 = -\dot{p}_4 \; ; \; \dot{p}_2 = -\dot{p}_3 \; ; \; \dot{p}_1 = \dot{p}_3 \; ; \; \dot{p}_2 = \dot{p}_4 \tag{26}$$

$$p_1 + p_4 = \int_k^\infty N(x, 0, 1)dx = \lambda; \qquad p_2 + p_3 = \int_{-\infty}^{-k} N(x, 0, 1)dx = \lambda.$$

If after making appropriate substitutions from (26) we set the right of (25) equal to zero we get

$$\frac{n_1 \dot{p}_1}{p_1} - \frac{n_2 \dot{p}_1}{\lambda - p_1} + \frac{n_3 \dot{p}_1}{p_1} - \frac{n_4 \dot{p}_1}{\lambda - p_1} = 0,$$

and since in general $\dot{p}_1 \neq 0$, the condition is that

$$\frac{n_1 + n_3}{n_2 + n_4} = \frac{p_1}{\lambda - p_1}. \tag{27}$$

Unless all four of the n_i are zero (which is unlikely for reasonable values of λ because n is large), it is possible to find a value of ρ which will make the right side of (27) equal to the ratio formed from the observations on the left, and the value of ρ so determined is the maximum likelihood estimate ρ under the restrictions we have imposed. In practice this equation may be solved by con-

sulting a table of the bivariate normal distribution—see for example K. Pearson [14]. Alternatively [27] may be solved by referring to Figure 3. Truman Kelley [9, 1939] has considered a closely related problem in connection with the validation of test items.

It may be inquired whether it would not be preferable to reduce the present design to a tetrachoric case by using only the cutting lines $x = 0$, $y = 0$. An investigation of the variance of $\tilde{\rho}$ reveals that such is not the case. We proceed to determine the asymptotic variance by means of the usual maximum likelihood technique. Differentiating (25) once more we have

$$\frac{d^2(\log g)}{d\rho^2} = \sum_{i=1}^{4} \frac{n_i(p_i \ddot{p}_i - \dot{p}_1^2)}{p_1^2}, \tag{28}$$

where $\ddot{p}_1 = \dfrac{d^2 p_1}{d\rho^2}$. We note that $E(n_i) = np_i$, therefore

$$E\left(\frac{d^2(\log g)}{d\rho^2}\right) = n\left(\sum_{i=1}^{4} \ddot{p}_i - \sum_{i=1}^{4} \frac{\dot{p}_1^2}{p_i}\right). \tag{29}$$

but since the derivative of a sum is equal to the sum of its derivatives, and $p_1 + p_4 = \lambda$, $p_2 + p_3 = \lambda$, the first sum in the square brackets vanishes. Suitable substitutions from (26) will reduce the second sum so that we get

$$-E\left[\frac{d^2(\log g)}{d\rho^2}\right] = \frac{2n\dot{p}_1^2\lambda}{p_1(\lambda - p_1)}. \tag{30}$$

Therefore asymptotically $\tilde{\rho}$ is normally distributed with variance

$$\sigma^2(\tilde{\rho}) = \frac{p_1(\lambda - p_1)}{2n\lambda\dot{p}_1^2}. \tag{31}$$

In general the optimum value (in the minimum variance sense) of λ which determines the cutting lines $x = \pm k$ will depend on the true value of ρ. To carry out the minimization process in general will require fairly extensive computations, which we feel would be justified. For the present we shall restrict ourselves to minimizing $\sigma^2(\tilde{\rho})$ for the case $\rho = 0$.

We have

$$\dot{p}_1 = \frac{1}{2\pi}\exp\left[-\tfrac{1}{2}k^2\right] = \frac{1}{\sqrt{2\pi}}f(k).$$

when $\rho = 0$, and $p_1 = \frac{1}{2}\lambda$. This gives

$$\sigma^2(\tilde{\rho} \mid \rho = 0) = \frac{\pi\lambda}{4n[f(k)]^2}. \tag{32}$$

We wish to minimize the expression on the right. We recall that a similar expression λ_1/f_1^2 was to be minimized in section 3 when the optimum pair of

observations for estimating the mean of a normal distribution was found. Using the previous results we have $\lambda \doteq .2702$, $k = .6121$; which gives us finally

$$\sigma^2_{\text{opt}}(\tilde{\rho} \mid \rho = 0) \doteq \frac{1.939}{n}. \tag{33}$$

To summarize: if a sample of size n is drawn from a normal bivariate population with known means a_x, a_y and variances σ_x^2 and σ_y^2, but unknown correlation ρ, the maximum likelihood estimate of ρ based on the number of observations falling in the four corners of the plane determined by the lines $x = a_x \pm k\sigma_x$, $y = a_y$ is found by solving for ρ the equation

$$\frac{n_1 + n_3}{n_1 + n_2 + n_3 + n_4} = \frac{p_1}{\lambda}.$$

where n_1 is the number of observations falling in the upper right, n_2 in the upper left, n_3 in the lower left, n_4 in the lower right hand corner, and p_i is the probability density in the region into which the n_i fall, $\lambda = p_1 + p_4$. The variance of this estimate $\tilde{\rho}$ is given by

$$\sigma^2(\tilde{\rho}) = \frac{p_1(\lambda - p_1)}{2n\lambda p_1^2},$$

which is minimized for $\rho = 0$ by setting $k = .6121$, $\lambda = .2702$, giving

$$\sigma^2_{\text{opt}}(\tilde{\rho} \mid \rho = 0) \doteq \frac{1.939}{n}.$$

On the other hand if the usual tetrachoric estimate is used with $x = 0$, $y = 0$ as the cutting lines we get $\sigma_T^2(\tilde{\rho} \mid \rho = 0) = \pi^2/4n$. The relative efficiency of the tetrachoric compared with the optimum statistic is therefore .787. The variance of the efficient estimate r given in (25) when $\rho = 0$ is $1/n$. Consequently the efficiency of our estimate $\tilde{\rho}$ compared to that of r is about .515 for the special case $\rho = 0$ under consideration. This means about twice as large a sample is required to get the same precision with $\tilde{\rho}$ as with r. Doubling the sample and using the cruder statistic $\tilde{\rho}$ may often be an economical procedure.

It may be surmised that a still better estimate of ρ could be constructed by employing four cutting lines, say $x = \pm k$, $y = \pm k$. The simplifications which we used to obtain the estimate $\tilde{\rho}$ no longer hold when we use this new construction. However, it is still possible to compute the minimum variance of the new estimate which we will call $\tilde{\rho}'$, for the special case $\rho = 0$. It again turns out that $k \doteq .6121$ minimizes and we get

$$\sigma^2_{\text{opt}}(\tilde{\rho}' \mid \rho = 0) \doteq \frac{1.52}{n}, \tag{34}$$

which makes the efficiency of $\tilde{\rho}'$ (compared with r) about .66 as compared with .515 for $\tilde{\rho}$. This suggests that if some very simple technique can be found for obtaining $\tilde{\rho}'$, $\tilde{\rho}'$ would be worth using. Unfortunately the author has not been able to construct a rapid way of finding $\tilde{\rho}'$.

5B. Estimation of ρ when the parameters are unknown. A more practical situation than the case treated in paragraph 5A, is the case in which all parameters of (20) are unknown. This case will be treated by means of order statistics. We construct an order statistic analogue of the estimate $\tilde{\rho}$ which we will call $\tilde{\rho}^*$. In general the procedure will be as follows: Each of the N observations in the sample has an x coordinate and a y coordinate

i) order the observations with respect to the x coordinate;

ii) discard all observations except the n with the largest x coordinates called the *right* set and the n with the smallest x coordinates called the *left* set, retaining, therefore, $2n$ observations;

iii) order the pooled $2n$ observations with respect to the y coordinate;

iv) break the $2n$ observations into two sets of n observations each; the *upper* set containing the n observations with the greatest y coordinates, and the *lower* set containing the n observations with the smallest y coordinates;

v) reorder the *upper* set of observations with respect to the x coordinate; the n observations will be divided into those whose x coordinates belong to the *right* set and those whose x coordinates belong to the *left* set;

vi) the estimate $\tilde{\rho}^*$ will be obtained by solving the equation

$$\frac{n_1^*}{n - n_1^*} = \frac{p_1^*}{\lambda_1^* - p_1^*}. \tag{35}$$

where n_1^* is the number of observations in the *upper* set which are also numbers of the *right* set and p_1^* is $\int_0^\infty \int_{k^*}^\infty f(x, y)dx\,dy$, while $f(x, y)$ is the bivariate normal (20) with $\sigma_x = \sigma_y = 1$, $a = b = 0$, and $\int_{k^*}^\infty N(x, 0, 1)\,dx = \frac{n}{N} = \lambda_1^*$.

Figure 2 represents graphically the construction described above for a scatter diagram composed of 25 observations. Of course the number 25 is only for purposes of illustration, as the method is only proposed for use with large samples.

The procedure of ordering the x's and choosing the right and left sets of observations is analogous to cutting the bivariate distribution by the two lines $x = \pm k$ as described in paragraph 5A, indeed $x = x_{n+1}$ and $x = x_{N-n}$ are the corresponding lines, but they vary from sample to sample. To continue the analogy, ordering the remaining observations with respect to y and dividing them into *upper* and *lower* sets of equal size is like cutting the plane with the line $y = 0$. Finally formula (35) is analogous to formula (27). Another similar change is that where formerly we had among relations (26) the equalities $p_1 =$

p_3 , $p_2 = p_4$, we now have the corresponding relations amongst the number of observations in the four corners of the plane, namely $n_1^* = n_3^*$, $n_2^* = n_4^*$ which

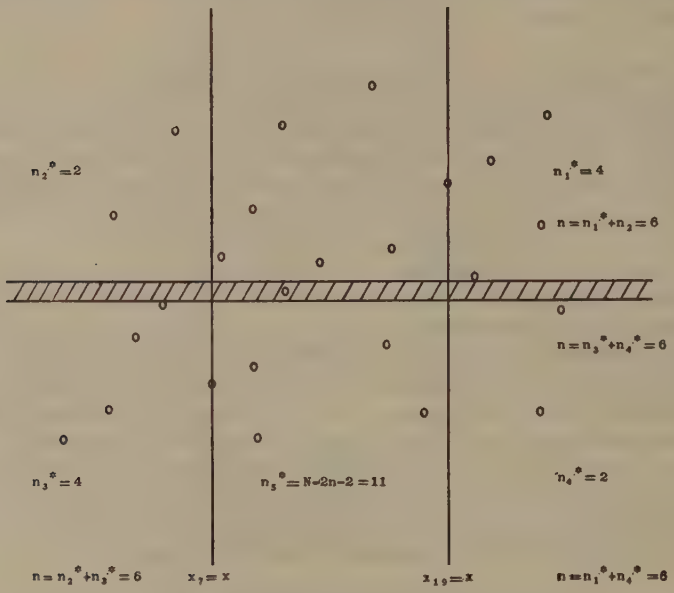

FIG. 2. DIAGRAM OF THE CONSTRUCTION DESCRIBED IN PARAGRAPH 5B ON THE BASIS OF 25 OBSERVATIONS $n = 6$

can readily be seen by inspection of the fourfold table we have constructed below (omitting all reference to $N - 2n$ pairs of observations we have discarded).

	Left set	Right set	Totals
Upper set	n_2^*	n_1^*	n
Lower set	n_3^*	n_4^*	n
Totals	n	n	$2n$

We have dwelt at length upon the analogy between the two constructions because one of the principal difficulties in working with order statistics is to design a mathematically workable model. The author has found it fruitful when constructing systematic statistics to study a workable analogy which does not involve the order statistics directly, and then to build upon correspondences such as those described.

Some may not wish to read further in this paragraph when they are informed that asymptotically the variance of $\bar{\rho}^*$ is essentially the same as that of $\bar{\rho}$. They should proceed to page 404. For the others we proceed to the demonstration.

Suppose we draw a sample of N pairs of observations (x_i', y_i') from the bivariate normal (20). If we discard from these all pairs except those with the n largest x_i and the n smallest x_i, we are left with the right set and the left set. We shall need the joint distribution of x_{n+1} and x_{N-n}

$$J(x_{n+1}, x_{N-n}) = \frac{N!}{(n!)^2(N-2n-2)!}\left(\int_{-\infty}^{x_{n+1}} g(x)\,dx\right)^n \left(\int_{x_{n+1}}^{x_{N-n}} g(x)\,dx\right)^{N-2n-2}\left(\int_{x_{N-n}}^{\infty} g(x)\,dx\right)^n g(x_{n+1}))g(x_{N-n}). \tag{36}$$

where $g(x)$ is the marginal distribution of x obtained from (20), $N(x, a, \sigma_x^2)$. We assume x_{n+1}, x_{N-n} satisfy Condition 1. Considering x_{n+1}, x_{N-n} as fixed and given for the moment we wish to look at the distribution of the y coordinates. We may consider the y coordinates of the observations in the right set as drawn from the distribution of y

$$\varphi'(y) = \frac{\int_{x_{N-n}}^{\infty} f(x, y)\,dx}{\int_{-\infty}^{\infty}\int_{x_{N-n}}^{\infty} f(x, y)\,dx\,dy} = \frac{\int_{x_{N-n}}^{\infty} f(x, y)\,dx}{\int_{x_{N-n}}^{\infty} g(x)\,dx}.$$

Similarly the y coordinates of the observations belonging to the left set may be considered as independently drawn from

$$\psi'(y) = \frac{\int_{-\infty}^{x_{n+1}} f(x, y)\,dx}{\int_{-\infty}^{\infty}\int_{-\infty}^{x_{n+1}} f(x, y)\,dx\,dy} = \frac{\int_{-\infty}^{x_{n+1}} f(x, y)\,dx}{\int_{-\infty}^{x_{n+1}} g(x)\,dx}.$$

To prevent confusion, in considering the y order statistics of the two sets, we shall designate those of the observations which are members of the right set by $u_1, u_2, \cdots, u_n$; while those observations belonging to the left set will have their ordered y coordinates designated $v_1, v_2, \cdots, v_n$. Of course the u's and v's separately satisfy an order relation like that given in (1).

The first question we answer is: given x_{n+1}, x_{N-n}, what is the probability that when we collate the u's and v's and split the observations into the upper set and lower set (see iv). there will be exactly c observations in the lower set whose y coordinates are designated by u's? In other words what is the probability that exactly c members of the lower set belong to the right set? An example for small values of n may clarify the problem. Suppose $n = 4$, and we observe $u_1 < v_1 < v_2 < v_3 < u_2 < u_3 < v_4 < u_4$; the y coordinates of the lower set of observations are u_1, v_1, v_2, v_3, and only the observation with u_1 for its y coordinate belongs to the right set, so for this case $c = 1$. To return

to our general problem, the probability that there are exactly c observations which are members of both the right set and the lower set is

$$P(c \mid x_{n+1}, x_{N-n}) = 1 - p(v_{n-c} > u_{c+1}) - p(u_c > v_{n-c+1}), \tag{37}$$

where $p(w > z)$ is the probability that w is greater than z. Now writing $\varphi(z) = \int_{-\infty}^{z} \varphi'(t)dt$; $\psi(z) = \int_{-\infty}^{z} \psi'(t)dt$ we may rewrite (37) as

$$\begin{aligned} &P(c \mid x_{n+1}, x_{N-n}) \\ &= 1 - \frac{n!}{c!(n-c-1)!} \int_{u_{c+1}}^{\infty} [\psi(v_{n-c})]^{n-c-1}[1 - \psi(v_{n-c})]^{c} \psi'(v_{n-c})\, dv_{n-c} \\ &- \frac{n!}{(n-c)!(c-1)!} \int_{-\infty}^{u_c} [\psi(v_{n-c+1})]^{n-c}[1 - \psi(n_{n-c+1})]^{c-1} \psi'(v_{n-c+1})\, dv_{n-c+1}. \end{aligned} \tag{38}$$

After integrating the first integral of (38) by parts and simplifying we can rewrite (38) as

$$\begin{aligned} P(c \mid x_{n+1}, x_{N-n}) &= \frac{n!}{c!(n-c)!} [\psi(u_{c+1})]^{n-c}[1 - \psi(u_{c+1})]^{c} \\ &+ \frac{n!}{(n-c)!(c-1)!} \int_{\psi(u_c)}^{\psi(u_{c+1})} \alpha^{n-c}(1-\alpha)^{c-1}\, d\alpha. \end{aligned} \tag{39}$$

We approximate the integral term of (39) by

$$\frac{n!}{(n-c)!(c-1)!} [\psi(u_{c+1}) - \psi(u_c)][\psi(u_{c+1})]^{n-c}[1 - \psi(u_{c+1})]^{c-1}$$

which leads us to the approximation

$$\begin{aligned} &P(c \mid x_{n+1}, x_{N-n}) \\ &= \frac{n!}{(n-c)!c!} [\psi(u_{c+1})]^{n-c}[1 - \psi(u_{c+1})]^{c-1}[1 + (c-1)\psi(u_{c+1}) - c\psi(u_c)]. \end{aligned} \tag{40}$$

The joint distribution of u_c, u_{c+1} is given by

$$\begin{aligned} &Q(u_c, u_{c+1} \mid x_{N-n}) \\ &= \frac{n!}{(c-1)!(n-c-1)} \varphi(u_c)^{c-1}(1 - \varphi(u_{c+1}))^{n-c-1} \varphi'(u_c)\varphi'(u_{c+1}). \end{aligned} \tag{41}$$

Next we multiply P as given by (40) by Q from (41) and integrate out u_c. This gives us except for terms of $O\left(\frac{1}{n}\right)$ and higher

$$\begin{aligned} &\frac{n!n!}{c!(n-c-1)!c!(n-c)!} [\varphi(u_{c+1})]^{c} \\ &\cdot [1 - \varphi(u_{c+1})]^{n-c-1}[\psi(u_{c+1})]^{n-c}[1 - \psi(u_{c+1})]^{c} \varphi'(u_{c+1}). \end{aligned} \tag{42}$$

When expression (42) is multiplied by (36), we finally get the approximate joint distribution of c, x_{n+1}, x_{N-n}, u_{c+1}.

Before proceeding further we let

$$\varphi(u_{c+1}) = \frac{\int_{-\infty}^{u_{c+1}} \int_{x_{N-n}}^{\infty} f(x, y)\, dx\, dy}{1 - \lambda_2^*} = \frac{p_4^*}{1 - \lambda_2^*},$$

$$\psi(u_{c+1}) = \frac{\int_{-\infty}^{u_{c+1}} \int_{-\infty}^{x_{n+1}} f(x, y)\, dx\, dy}{\lambda_1^*} = \frac{p_3^*}{\lambda_1^*}, \tag{43}$$

where $\lambda_1^* = \int_{-\infty}^{x_{n+1}} g(x)dx$, $\lambda_2^* = \int_{-\infty}^{x_{N-n}} g(x)dx$. If we also let $p_1^* = 1 - \lambda_2^* - p_4^*$, $p_2^* = \lambda_1^* - p_3^*$ we can write

$$R(c,\ x_{n+1},\ x_{N-n},\ u_{c+1}) = K(\lambda_2^* - \lambda_1^*)^{N-2n-2} p_1^{*n-c-1} p_2^{*c} p_3^{*n-c} p_4^{*c} p_4^{*\prime} \lambda_1^{*\prime} \lambda_2^{*\prime}, \tag{44}$$

where the primes indicate derivatives of p_4^*, λ_1^*, λ_2^* with respect to the appropriate suppressed variables, u_{c+1}, x_{n+1}, x_{N-n}, respectively.

We now proceed to the maximum likelihood estimate of ρ. We take the logarithm of (44) and then take partial derivatives with respect to a the mean of x, b the mean of y, and ρ the correlation coefficient. After equating these partial derivatives to zero we have the following three maximum likelihood equations which must be solved simultaneously to obtain the estimates $\hat{a}^*$, $\hat{b}^*$, and $\hat{\rho}^*$:

$$\frac{1}{N}\left[\frac{N - 2n - 2}{\lambda_2^* - \lambda_1^*} \frac{\partial(\lambda_2^* - \lambda_1^*)}{\partial a} + \frac{n - c - 1}{p_1^*} \frac{\partial p_1^*}{\partial a} + \frac{c}{p_2^*} \frac{\partial p_2^*}{\partial a} + \frac{n - c}{p_3^*} \frac{\partial p_3^*}{\partial a} + \frac{c}{p_4^*} \frac{\partial p_4^*}{\partial a}\right] = 0, \tag{45}$$

$$\frac{1}{N}\left[\frac{n - c - 1}{p_1^*} \frac{\partial p_1^*}{\partial b} + \frac{c}{p_2^*} \frac{\partial p_2^*}{\partial b} + \frac{n - c}{p_3^*} \frac{\partial p_3^*}{\partial b} + \frac{c}{p_4^*} \frac{\partial p_4^*}{\partial b}\right] = 0, \tag{46}$$

$$\frac{1}{N}\left[\frac{n - c - 1}{p_1^*} \frac{\partial p_1^*}{\partial \rho} + \frac{c}{p_2^*} \frac{\partial p_2^*}{\partial \rho} + \frac{n - c}{p_3^*} \frac{\partial p_3^*}{\partial \rho} + \frac{c}{p_4^*} \frac{\partial p_4^*}{\partial \rho}\right] = 0. \tag{47}$$

where terms $O\left(\frac{1}{N}\right)$ have been neglected. Equations (45) and (46) are satisfied, again except for terms $O\left(\frac{1}{N}\right)$, when $\hat{a}^* = \frac{1}{2}(x_{n+1} + x_{N-n})$, $\hat{b}^* = u_{c+1}$. Using this information we examine (47) and find it satisfied when

$$\frac{n - c}{c} = \frac{p_1^*}{\lambda_1^* - p_1^*}, \tag{48}$$

which is directly analogous to equation (27), and is the form promised in (35), if $n_1^* = n - c$. The estimate $\hat{\rho}^*$ is obtained by solving (48) for ρ, where $p_1^* =$

$\int_0^\infty \int_{k^*}^\infty f(x, y)\, dx\, dy$, and $f(x, y)$ is given by (20) with variances equal unity and means equal zero, and $\int_{k^*}^\infty g(x)\, dx = \lambda_1^* = 1 - \lambda_2^* = n/N$.

We shall not go through the derivation of $\sigma^2(\tilde{\rho}^*)$ here. The usual maximum likelihood technique may be used. It turns out that the covariances between $\hat{a}^*$ and $\tilde{\rho}^*$ and between $\hat{b}^*$ and $\tilde{\rho}^*$ are $O\left(\frac{1}{N^2}\right)$. Neglecting such terms we find that the variance is

$$\sigma^2(\tilde{\rho}^*) = \frac{p_1^*(\lambda_1^* - p_1^*)}{2N\lambda_1^* p_1^{*2}}. \tag{49}$$

To summarize: if a sample of size N is drawn from a normal bivariate population with unknown parameters, the maximum likelihood estimate of ρ based on the $2n$ observations composed of those observations with the n largest x coordinates and the n smallest x coordinates, may be obtained by solving for ρ the equation

$$\frac{n - c}{n} = \frac{p_1^*}{\lambda^*},$$

where $\frac{1}{2} > \lambda^* = n/N > 0$, $p_1^* = \int_0^\infty \int_{k^*}^\infty f(x, y \mid \sigma_x = 1, a_x = a_y = 0)\, dx\, dy$, $\int_{k^*}^\infty N(x, 0, 1)\, dx = \lambda^*$, *and* $n - c$ *is the number of the* $2n$ *observations with largest y coordinates, which also have largest x coordinates. The variance of this estimate* $\tilde{\rho}^*$ *is given by*

$$\sigma^2(\tilde{\rho}^*) = \frac{p_1^*(\lambda_1^* - p_1^*)}{2N\lambda^* p_1^{*2}},$$

and for $\rho = 0$ *the variance is minimized by choosing* $\lambda^* = .2702$, *that is by choosing that 27 per cent of the observations with largest x coordinates, and that 27 per cent with smallest x coordinates, and for this value of* λ^*

$$\sigma^2_{\text{opt}}(\tilde{\rho}^* \mid \rho = 0) \doteq \frac{1.939}{N}.$$

Equation (49) is of course exactly analogous to the expression given in (31) for the case of known means and variances. Therefore if the variance minimization problem is solved in general for the case of paragraph 5A, the large sample solution of the problem for unknown means and variances will also be solved.

Figure 3 may be used to obtain the estimates $\tilde{\rho}$ or $\tilde{\rho}^*$ in case the methods of paragraphs 5A or 5B are used. Essentially the figure solves equations (27) and (48). The procedure for the problem of paragraph 5A is

i) when $n_1 + n_3 > n_2 + n_4$ evaluate the ratio $\frac{n_1 + n_3}{n_1 + n_2 + n_3 + n_4} = x_0$ and

find the intersection of the line $x = x_0$ with the curve for the particular λ being used;

ii) through the point of intersection of the vertical line $x_0 = x$ and the λ curve draw a horizontal line;

iii) the value of $\tilde{\rho}$ is indicated on the vertical axis at the point of intersection of the horizontal line and the vertical axis;

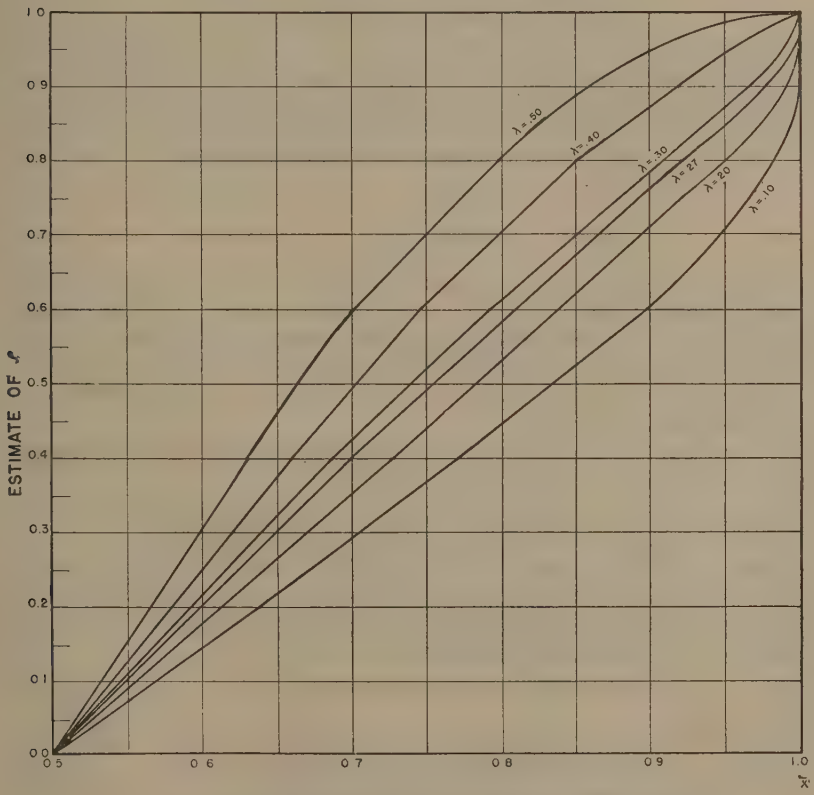

FIG. 3. CURVES FOR ESTIMATING THE CORRELATION COEFFICIENT ρ

iv) when $n_1 + n_3 < n_2 + n_4$ use the ratio $x_0 = \dfrac{n_2 + n_4}{n_1 + n_2 + n_3 + n_4}$ and follow the same procedure, $\tilde{\rho}$ will be the negative of the number appearing on the vertical axis.

Example. Suppose a sample of 1000 is drawn from a normal bivariate population for which the mean of x is a, and the mean of y is b, and the variance of x is σ_x^2, all three parameters known (it is not necessary to know σ_y^2). The xy

plane is cut by the three lines $x = a \pm k\sigma_x$, $y = a$, where, say, $k = .612$, so that $\lambda = .27$. Suppose we find the observations are distributed as follows:

in the upper right-hand corner: $160 = n_1$

in the lower left-hand corner: $170 = n_3$

in the upper left-hand corner: $110 = n_2$

in the lower right-hand corner: $110 = n_4$.

To estimate $\tilde{\rho}$ we set up $x_0 = (n_1 + n_3)/(n_1 + n_2 + n_3 + n_4) = 330/550 = .6$. Referring to Figure 3 we find that the estimate of ρ, $\tilde{\rho} = .20$.

In using Figure 3 for this case it is useful to know that for

$\lambda = .50$	$k = 0.000$	$\lambda = .27$	$k = 0.612$
$\lambda = .40$	$k = 0.253$	$\lambda = .20$	$k = 0.841$
$\lambda = .30$	$k = 0.524$	$\lambda = .10$	$k = 1.282$

If the means and variances of the variables are unknown, we may use the method of paragraph 5B:

i) when $n - c > c$ evaluate the ratio $(n - c)/n = x_0$, and find the intersection of the line $x = x_0$ with the curve for the particular λ_1 being used;

ii) through the point of intersection of the vertical line $x_0 = x$ and the λ_1 curve draw a horizontal line;

iii) the value of $\tilde{\rho}^*$ is indicated on the vertical axis at the point of intersection of the horizontal line and the vertical axis;

iv) when $n - c < c$, use the ratio $c/n = x_0$ and follow the same procedure, $\tilde{\rho}^*$ will be the negative of the number appearing on the vertical axis.

Example: Suppose a sample of 1000 is drawn from a normal bivariate population with all parameters unknown. Suppose we set $n = 200$, and follow the procedure given in paragraph 5B of this section, and suppose we find the observations are distributed as follows:

in the upper right-hand corner: $50 = n - c$

then of course

in the lower left-hand corner: $50 = n - c$

in the upper left-hand corner: $150 = c$

in the lower right-hand corner: $150 = c$

The estimate this time is clearly negative, so we set $x_0 = c/n = 150/200 = .75$. Referring to Figure 3 we find using the curve corresponding to $\lambda = .20$ that the estimate of ρ, $\tilde{\rho} = -.44$.

5C. The use of averages for estimating ρ when the variance ratio is known. Nair and Shrivastava [12, 1942] have considered the use of means for estimating

regression coefficients when one observation is taken at each of n equally spaced fixed variates, x_i ($i = 1, 2, \cdots, n$), and y is normally distributed. Their procedure was essentially to consider the ordered fixed variates, and to discard a group of observations in the interior, much as we discarded the set of observavations whose x coordinates were $x_{n+1}, x_{n+2}, \cdots, x_{N-n}$ in paragraph 5B. The resulting estimates depended essentially on the averages of the y's on the right and left sets of observations, and on the averages of the fixed x's in the two sets.

In an unpublished manuscript George Brown has considered a problem even more closely related to the one considered in paragraph 5A. Suppose x and y normally distributed according to (20) with equal variances σ^2, and means equal to zero. (The ratio of variances must be known, equality is unnecessary.) Retain only those observations for which $|x_i| \geq k\sigma$, and from them form the statistic

$$\rho_B = \frac{\bar{y}_+ - \bar{y}_-}{\bar{x}_+ - \bar{x}_-}, \tag{50}$$

where $\bar{y}_+$ and $\bar{x}_+$ are the average of the n_1 x's and y's for which $x_i > k\sigma$ and $\bar{y}_-$ and $\bar{x}_-$ are similarly defined for the n_2 observations for which $x_i < -k\sigma$. Then ρ_B is an unbiased estimate of ρ. Regarding the x's as fixed variates it turns out that

$$\sigma^2(\rho_B) = \frac{(1 - \rho^2)\sigma^2}{(\bar{x}_+ - \bar{x}_-)^2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right). \tag{51}$$

If we approximate by substituting expected values for observed values (55) turns out to be $(1 - \rho^2)\sigma^2\lambda/2N[g(k)]^2$, where $\lambda = \int_{-\infty}^{-k} g(x)\,dx$, $g(x) = N(x, 0, 1)$. The value of k which minimizes this expression is our old friend $k = .6121$, which gives $\lambda = .2702$. Therefore for $\rho = 0$ and large samples, the minimum variance is approximately $1.23\,\sigma^2/N$, for an efficiency of about .81. The relative efficiency of the methods of paragraphs 5A and 5B are .635 compared with the present technique.

We presume that the analogous order statistics construction would produce much the same result. Our interest in the present technique is to supply an approximate answer to the question of what is to be gained by going from the counting technique proposed in paragraph 5B to the next level of computational difficulty—addition.

6. Acknowledgements. The author wishes to acknowledge the valuable help received from S. S. Wilks, under whose direction this work was done, and the many suggestions and constructive criticisms of J. W. Tukey. The author also wishes to acknowledge the debt to his wife, Virginia Mosteller, who prepared the manuscript and assisted in the preparation of the tables and figures.

REFERENCES

[1] G. W. Brown and J. W. Tukey, "Some distributions of sample means," *Annals of Math. Stat.*, Vol. 17 (1946), p. 1.

[2] J. H. Curtiss, "A note on the theory of moment generating functions," *Annals of Math. Stat.*, Vol. 13 (1942), p. 430.

[3] R. A. Fisher and L. H. C. Tippett, "Limiting forms of the frequency distribution of the largest or smallest member of a sample," *Proc. Camb. Phil. Soc.*, Vol. 24 (1928), p. 180.

[4] R. A. Fisher and F. Yates, *Statistical Tables*, Oliver and Boyd, London, 1943.

[5] E. J. Gumbel, "Les valeurs extremes des distribution statistiques," *Annales de l'Institute Henri Poincare*, Vol. 5 (1934), p. 115.

[6] E. J. Gumbel, "Statische Theorie der grössten Werte," *Zeitschrift für schweizerische Statistik und Volkswirtschaft*," Vol. 75, part 2 (1939), p. 250.

[7] H. O. Hartley, "The probability integral of the range in samples of n observations from a normal population," *Biometrika*, Vol. 32 (1942), p. 301.

[8] D. Jackson, "Note on the median of a set of numbers," *Bull. Amer. Math. Soc.*, Vol. 27 (1921), p. 160.

[9] T. Kelley, "The selection of upper and lower groups for the validation of test items," *Jour. Educ. Psych.*, Vol. 30 (1939), p. 17.

[10] M. G. Kendall, *The Advanced Theory of Statistics*, J. B. Lippincott Co., 1943.

[11] J. F. Kenney, *Mathematics of Statistics*, Part II, D. Van Nostrand Co., Inc., 1939, Chap. VII.

[12] K. R. Nair and M. P. Shrivastava, "On a simple method of curve fitting," *Sankhya*, Vol. 6, part 2 (1942), p. 121.

[13] K. Pearson, "On the probable errors of frequency constants, Part III," *Biometrika*, Vol. 13 (1920), p. 113.

[14] K. Pearson (Editor), *Tables for Statisticians and Biometricians*, Part II, 1931, p. 78, Table VIII.

[15] N. Smirnoff, "Über die Verteilung des allgemeinen Gliedes in der Variationsreihe," *Metron*, Vol. 12 (1935), p. 59.

[16] N. Smirnoff, "Sur la dependance des membres d'un series de variations," *Bull. Univ. État Moscou, Series Int., Sect. A., Math. et Mécan.*, Vol. 1, fasc. 4, p. 1.

[17] L. H. C. Tippett, "On the extreme individuals and the range of samples taken from normal population," *Biometrika*, Vol. 17 (1925), p. 364.

[18] S. S. Wilks, *Mathematical Statistics*, Princeton University Press, Princeton, N. J., 1943.